

*Focus***Amino Acids as RNA Ligands: A Direct-RNA-Template Theory for the Code's Origin****Michael Yarus**

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309-0347, USA

Received: 22 December 1997 / Accepted: 13 February 1998

**Abstract.** Numerous RNA binding sites for specific amino acids are now known, coming predominantly from selection-amplification experiments. These sites are chemically discriminating despite being predominantly small, simple RNA structures: internal and bulge loops. Recent studies of sites for hydrophobic side chains suggest that there are other generalizable structural features which recur in hydrophobic RNA sites. Further, sites for hydrophobic side chains can contain codons for the bound amino acid, as has also long been known for the polar amino acid arginine. Such findings are comprehensively reviewed, and the implications for the origin of coded peptide synthesis are considered. An origins hypothesis which accommodates all the data, DRT (direct RNA templating), is formulated.

**Key words:** Selection-amplification — Genetic code — Affinity selection — RNA structure — Internal loop

**Introduction**

Small RNAs can be folded to yield binding sites for varied amino acids. Such RNA binding sites not only distinguish similar side chains, but can be quite selective for L-amino acids over the D-enantiomer. Given such a portfolio of information about amino acid:RNA affinity, to what uses can it be put? One possibility is that such individual RNA sites can be the building blocks from which biological protein–RNA interfaces are con-


structed. For example, study of arginine:RNA complexes has played a substantial role in the subject of regulatory arginine-rich peptide:RNA affinities (Tan and Frankel 1995) such as that in retroviral transcriptional control via Tat:TAR RNA complexes.

However, such binding interactions necessarily associate particular RNA sequences (within the sites) with particular amino acids. This parallels the logic of the genetic code, which also associates RNA sequences with individual amino acids. It is worth inquiring whether there is any connection. Are amino acid:RNA associations within binding sites anything like amino acid:RNA associations preserved in the code? Recently, this possibility has been briefly treated, with varying conclusions about the application of binding-site data to this problem (Hirao and Ellington 1995; Cedergren and Miramontes 1996). However, a comprehensive treatment does not exist and seems due. Below I discuss all published, characterized, specific interactions between free amino acids and RNA sequences, to see if they can sustain a rationalization of the genetic code in terms of presently demonstrable RNA chemistry.

**Sites Considered**

There are three pieces of work that are not discussed in detail below, however, because the present analysis both calls for RNA sites directed at the amino acids alone (not at sites which also include other features) and, in addition, requires sites whose nucleotide sequences are known. A reader interested in a more complete survey might refer to these three omissions: in one case, an RNA

triplet	no.	freq.
AGA	319	0.71
AGG	11	0.025
AGU	0	0.0
AGC	5	0.011
CGA	94	0.21
CGG	1	0.002
CGU	0	0.0
CGC	17	0.038


  
 N'GNNN group I helix P7  
 N\_CNNNN

**Fig. 1.** Sequences of the triplets at the G/arginine site in 447 sequenced group I RNAs. Nucleotides 263 264 265 (*Tetrahymena* numbering) are shown. N, any nucleotide; N', any nucleotide, but complementary to N; Y, U or C; R, A or G; M, C or A; H, A or U or C.

was selected for affinity to tryptophan–agarose, but the product could not be shown to bind the free amino acid (Famulok and Szostak 1992). Probably RNA elements that bind the agarose matrix were included in the site. Second, DNA sites for arginine were selected (Harada and Frankel, 1995), but the change from RNA to DNA makes relevance to the genetic code arguable. Finally, phe/trp-binding RNAs were selected (Zinnen and Yarus 1995) but the binding sites have not been located within these RNAs, and they are therefore not useful here. The remaining experiments are seven selection-amplification (Ellington and Szostak 1990; Robertson and Joyce 1990; Tuerk and Gold 1990) experiments which isolated RNAs with an affinity for amino acid-containing columns and one natural-site amino acid (Yarus 1988) whose discovery introduced the possibility of specific amino acid: RNA association, with which we begin (Fig. 1).

## A Natural Example

The group I RNAs are catalytic introns. These frequently catalyze their own excision and splicing from a precursor RNA (e.g., Golden and Cech 1996). This process begins with catalysis of attack by a free molecule of the splicing cofactor, guanosine or G nucleotide, at the 5' exon–intron junction. Within the group I self-splicing RNAs there is a broadly conserved (Hicke et al. 1989) site for arginine, which can be moderately avid ( $K_D = 400 \mu M$ ), quite selective among the standard 20 amino acids (160:1 against L-lysine), and up to 10:1 selective (L:D) against D-arginine (Yarus and Majerfeld 1992). This arginine site exists within the guanosine splicing cosubstrate site because of a resemblance between the stacking (Yarus and Majerfeld 1992) and, particularly, the hydrogen-bonding (Yarus 1988) patterns of guanine and arginine. In accord with binding to the same groups, substitution of nucleotides within this RNA site alters the activities of guanosine and arginine in parallel (Yarus and Majerfeld 1992; Yarus et al. 1991a). Since the discovery of the

locus of the splicing cosubstrate site (Michel et al. 1991), it has been recognized (Yarus and Christian 1991) that while the nucleotides immediately at the guanosine/arginine site vary, they are virtually always triplets (Yarus et al. 1991b) which correspond to arginine codons. Thus free arginine binds to an RNA site containing its own coding triplets.

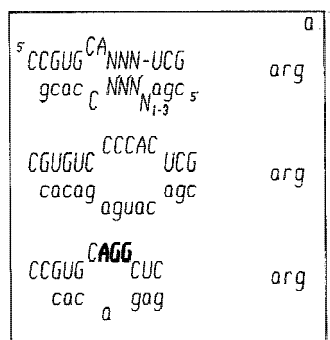
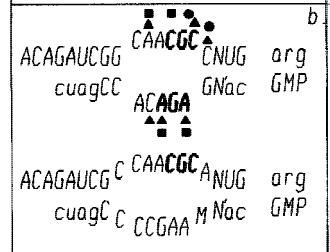
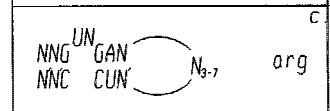
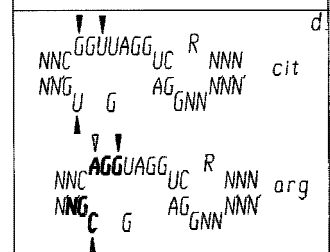
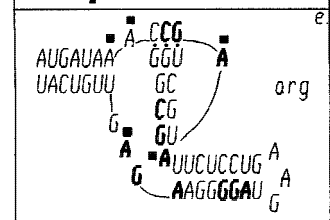
The group I RNA active-center sequence is updated in Fig. 1. The figure shows the triplets at the presumptive G/arg sites within the active-site P7 helix of 447 currently known group I RNA sequences from all phylogenetic groups. Only the nucleotides along one side of the helix are active in the binding site (Yarus et al. 1991a), as shown in Fig. 1. Canonical arginine codons are AGR and CGN; about 99% of 447 such active-site group I triplets are arginine codons, and five of six of the modern arginine codons have been observed (Fig. 1). The extreme conservation of this pattern makes it likely that arginine coding triplets which bind arginine are at least as old as the group I active center. Group I RNAs may be ancient (Shub 1991), perhaps ancient enough to be the progenitor of the code for arginine. A more comprehensive review of this site is available (Yarus 1993).

The detection of what could be a molecular fossil embodying the genetic code for arginine led to a more general search for coding sequences in amino acid-binding sites. Selection-amplification (Tuerk and Gold 1990; Robertson and Joyce 1990; Ellington and Szostak 1990) supplied a general technique for isolation of such RNAs. This procedure typically being with  $10^{14}$ – $10^{15}$  RNAs with different randomized sequences, derived by transcription of random-sequence DNA. The tiny minority of amino acid-binding RNAs in this population may be purified by affinity selection among these molecules using retardation on an amino acid-containing matrix and (usually) elution by free amino acid. Selected molecules can be amplified by conversion to cDNA and subsequent PCR. Transcription and repetition of this cyclic procedure finally yield novel RNAs with amino acid binding sites.

## A Review of Sites from Selection-Amplification

Figures 2 and 3 show the predominant RNA structures which met such affinity selections for amino acids. In each case the binding site was defined as closely as possible, so that these drawings generally contain fewer nucleotides than their originals. Conservations among molecules, single-ended truncation experiments to find minimal active structures (boundaries), remutagenesis, modification-interference, and other types of information have been used wherever possible to concentrate attention on the region of the RNA nearest the amino acid.

At first glance, the structures themselves appear similar in one sense. Save for one example, all amino acid

a	SELECTION	PREVALENCE	PROPERTIES
	Sephacrose-thiopropyl-cys-arg	60%	$K_D = 1 \text{ mM}$ for free arginine
	arg & GMP elution	55%	$K_D = 4 \text{ mM}$ for free arginine $K_D = 110 \mu\text{M}$ for free GMP
	Agarose-arg	41%	$K_D = 2 - 4 \text{ mM}$
	Agarose-cit	100%	$K_D = 65 \mu\text{M}$ for free citrulline
	Agarose-arginine	26%	$K_D = 0.33 \mu\text{M}$ for free arginine

**Fig. 2.** Structures and sequences of RNA sites selected to bind arginine. *Lowercase letters*, fixed nucleotide; *capitalized letters*, randomized nucleotide; *boldface*, arginine coding triplet whose identity is unforced. *Filled squares*, reactivity to DMS altered by arginine; *filled circles*, phosphate interference with arginine binding; *filled triangles*, base interference with binding to arginine column; *arrowheads*, base in proximity to bound amino acid side chain.

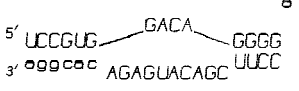
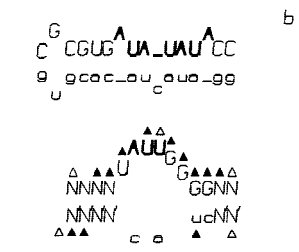
binding sites are within simple asymmetric internal or bulge loops, containing 1 to 10 nucleotides on each strand. This comparison extends even to the natural site for arginine in the group I active center. Thus small asymmetric loops are able to present chemically varied polar and hydrophobic surfaces despite their size and structural simplicity. This apparent potential for RNA sites of similar overall structures but specific for varied amino acids is discussed again below.

Within these structures, coding triplets are of particular interest. In Fig. 2, arginine triplets within arginine binding sites, like those in the group I active center (Fig. 1), are boldfaced for special consideration. But not all such triplets are marked. To be singled out, the identity of a triplet must be acquired during the selection, and must not be forced by trivial aspects of the experimental design. For example, a boldfaced triplet must occur in nucleotides originally randomized rather than sequences

initially fixed, unless subsequent randomization showed that the fixed sequences had become essential for site activity. As another example, triplets are not boldfaced if their only apparent role is as Watson-Crick pairs to initially fixed sequences. To help make such distinctions, nucleotides originally fixed are in lowercase letters, and nucleotides which were randomized in capitals.

### Sites Taken One by One

Now for a survey of results of selection: Fig. 2a shows the three most frequent RNAs found to bind arginine when affinity for Sepharose-arginine and elution by free L-arginine in the chromatographic buffer were used for selection. Boundaries and comparison of similar active sequences were used to assign the minimal structures shown. The molecules shown and their close congeners

	SELECTION	PREVALENCE	PROPERTIES
	Agarose-val valinamide elution 25 randomized nt 3 x 10 <sup>13</sup> seq	28%	K <sub>D</sub> = 12 mM
	Sepharose-ile isoleucine elution 50 randomized nt 2 x 10 <sup>13</sup> seq	18%  14%	K <sub>D</sub> = 10 mM  K <sub>D</sub> = 0.4 mM

**Fig. 3.** Structures and sequences of RNA binding sites for aliphatic amino acids. Conventions as in Fig. 2 except that *open triangles* indicate no interference with binding and elution by arginine.

account for 80% of the sequences in the final pool. The third most prevalent (third row), at 1 of 12 total pool sequences, contains a triplet corresponding to an arginine codon (boldface). Though the sequence of this triplet varied, it changed to AGA, another arginine codon (Connell et al. 1993).

Figure 2b shows two related sequences which are the results of a more complex selection, in which affinity and affinity elution occurred on alternating GMP and arginine columns. The plan of the selection (Connell and Yarus 1994) was to isolate sites which could be compared to the group I RNA, which binds both a nucleoside and an amino acid. In addition, determination of the minimal number of nucleotide changes required to alter emphasis between these specificities was a goal. A single Y-to-A transversion in the site appeared to cause a 100-fold change in the relative binding constants for GMP and arginine, indicating that this site withstands the single mutations needed to evolve a new specificity. Boundaries, synthesis of truncated RNAs, and the chemical modification studies shown make the neighborhood of the amino acid site apparent. In Fig. 2b squares appear above nucleotides protected from DMS by arginine, circles appear by phosphates protected, and triangles appear where DMS reaction interferes with affinity for the arginine column. The tight constellation of these functional nucleotides makes clear that the arginine site is close to two arginine triplets, CGC (above) and AGA (lower strand). This structure comprised the majority of sequences that met the initial selection. The lower structure, containing only the upper-strand triplet, is derived in parallel with the upper one by remutagenesis and alternating reselection.

Figure 2c shows the consensus sequence from the only selection for a property other than affinity elution by arginine. Here the ability to survive elevated NaCl elution from arg-agarose was selected. The purpose was to focus on the partially electrostatic interaction between the guanidinium group of the arginine side chain and the nucleotide phosphate. Such an interaction mimics the retroviral Tat-TAR RNA interaction (Tao and Frankel

1996). The recurring sequences found can often be summarized as variants of the natural HIV TAR hairpin, as shown. Only a few nucleotides are fixed, including only one run of three, which is not an arginine triplet.

Figure 2d summarizes a family of RNA sites (Famulok 1994; Burgstaller et al. 1995; Yang et al. 1996) for which refined NMR structural data (Yang et al. 1996) are available. In this case, the approach to arginine was indirect. An initial selection was for affinity elution with L-citrulline, an amino acid with a related structure, and only the RNA shown at the top in Fig. 2d met the selection (21 sequences). To see if the initial RNA specificity could be changed to arginine, and to determine how many sequence changes would be required, the initial isolate was resynthesized with 30% mutation at each position (10% of each nonparental nucleotide) and reselected by arginine elution from arg-agarose. The new structure at the bottom in Fig. 2d has three nucleotide changes and is highly specific for L-arginine.

The three changed nucleotides account for the change in side-chain specificity. All three contact the citrulline side chain specifically in the NMR structure of the original molecule (Fig. 2d; filled arrowheads at the top) (Yang et al. 1996). The three changes (Fig. 2d; filled arrowheads at the bottom) create two new arginine triplets and an arginine site. Nucleotides of both triplets are inside the van der Waal's radii of the arginine side chain. Triplet AGG makes contacts via all three of its bases. The A is packed against the amino acid (Fig. 2d; open arrowhead at the bottom), and the two G's each make multiple hydrogen bonds to arginine guanidinium. The C of CGN also makes multiple hydrogen bonds to the end of the arginine side chain. The contact between the new triplets and the amino acid is therefore intimate: as close as separate molecules usually get.

Finally in Fig. 2e, the most complex structure (an internal loop somewhat like the other sides but surrounding a self-contained pseudoknot) appears as the result of the most complex selection (Geiger et al. 1996). The structure shown was selected to resist several preliminary elutions from agarose-arginine, including elution

with 20 mM arginine at 23°C. It was collected after heat denaturation at 95°C using further arginine affinity elution. The binding site has exceptional affinity and DL-stereoselectivity (12,000-fold) for free L-arginine. The position of the arginine site is defined only by the five A's protected by arginine (from N-1 alkylation by DMS) scattered over the molecule. Remarkably, however, three of these A nucleotides are within arginine triplets (bold-face in Fig. 2e) and three of four such triplets in the site are involved. The fourth AGG triplet appears to be completely paired in a helical stem.

### Conclusions About Arginine:RNA Sites

These selections taken together make clear that there are a very large number of ways in which to generate specific RNA sites for arginine. No site predominant in one selection has ever been reisolated in another, independent experiment. Instead, each time the selection was changed, a new set of sequences, or usually several sets of new sequences, appeared. This undoubtedly reflects the stacking and H-bonding versatility of the guanidinium side chain of arginine, which is somewhat like a nucleotide in its planarity and H-bonding pattern.

Thus we are probably underestimating this variety. Selections are often progressively increased in rigor in successive cycles but are usually stopped short of the point at which only one sequence can satisfy the selective criterion. Thus in all selections for arginine affinity, the final pool contained 20–74% uncharacterized sequences, or sequences which at least go undescribed. Judging from experience, only a small minority of these will be nonfunctional noise, surviving selection for an irrelevant reason. Among this variety of rarer functional sequences, therefore, there will surely be the missing reisolated sequences from other selections, as well as new, yet undetected ways of folding sites for arginine.

With regard to triplets, in eight selected RNA binding sites for arginine, there are 11 arginine coding triplets which are arguably in, or close to, molecular contact with the amino acid itself and are not forced by the conditions of selection. Taken another way, five of the eight site structures contain such sequences. If the natural group I site is also considered, these statistics would be 12 triplets in 9 sites and 6 of 9 structures containing coding triplets. This includes structurally explicit cases such as in Fig. 2d, where both the arginine site and the two triplets contacted by arginine are simultaneously created by selection for arginine affinity (Yang et al. 1996). This kind of observation requires further interpretation, and I return to it below.

Because one might argue that the unique chemical character of arginine makes it a special case, it would be useful to have sites with other specificities to compare. Thus the next section.

### Sites for Aliphatic Side Chains

Figure 3 summarizes RNA sites selected for affinity for amino acids with aliphatic hydrophobic side chains: a site for L-valine (Majerfeld and Yarus 1994) and two more recently isolated sites which bind isoleucine (Majerfeld & Yarus, 1998).

Figure 3a shows a valine-specific site within an internal loop of 4 over 10 nucleotides. The binding reaction is notable for distinguishing aliphatic side chains of the same area but different shapes, and for its affinity per methylene group of about  $-1.5$  kcal/mol, not greatly different from many proteins which bind hydrophobic ligands. However, it contains no conserved coding triplets for valine among its nucleotides, though there is one (not shown) among the variants of the lower loop (Majerfeld and Yarus 1994).

Figure 3b shows the two predominant structures in a pool of RNAs selected to bind isoleucine (Majerfeld and Yarus 1998). The upper sequence is defined only by sequence conservations but contains the oligomer AUAUAUA (which can be read as overlapping isoleucine AUA codons). However, it was not analyzed extensively because it is not side chain selective.

The lower sequence in Fig. 3b contains an isoleucine binding site positioned within the 7 nucleotide over 2 loop shown. Binding has been localized by multiple criteria: by sequence conservation, by boundary experiment, by synthesis of truncated molecules and molecules of altered structure, and by the modification–interference data shown in the figure. The specific nucleotides of the loop, constrained by arbitrary flanking helices, create the site. Here an AUU triplet (an isoleucine codon) is a conserved feature of the isoleucine site, which also is L-stereoselective and selective among aliphatic amino acids. The discovery of apparently functional coding sequences within these sites for a new chemical class of side chains provided the impetus for this review.

### Conclusions About Hydrophobe:RNA Associations

RNAs provide selective sites not just for the intensely polar arginine side chain, but also for the chemically disparate aliphatic hydrocarbons of valine and isoleucine. Within the fewer number of sites for aliphatic side chains, asymmetric internal loops are again prominent structures, as they were for arginine. These simple molecules again make distinctions that might have once seemed outside the capability of even complicated RNA structures: in Fig. 3a L-valine is preferred to L-isoleucine, and in Fig. 3b this preference is reversed. Strings of G's associated with G:U appositions are prominent in both valine and isoleucine sites, and this is likely of functional significance (see below). Whatever the building blocks,



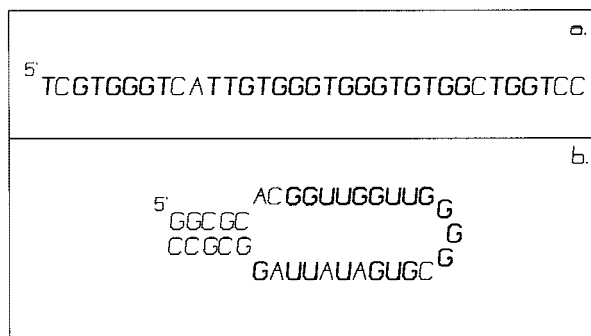
it is clear that hydrophobic elements must exist in RNA which can be assembled by a higher-order structure into sites of different shape and extent.

Some of these hydrophobic building blocks can be enumerated from structural data. In NMR and crystallographic structures of nucleic acid:peptide interfaces with a conspicuously hydrophobic character, purine base surfaces (Y Kim et al. 1993; JL Kim et al. 1993; Werner et al. 1995), the C'1H side of the sugar ring (Y Kim et al. 1993; JL Kim et al. 1993; Werner et al. 1995), and the minor groove edges of bases (Y Kim et al. 1993; JL Kim et al. 1993; Werner et al. 1995) apparently make stabilizing or minimally destabilizing contacts with aliphatic and aromatic amino acid sidechains.

Even more specifically, a structural model for ribonucleotide:isoleucine interaction exists. It is a variation of a theme just mentioned for DNA:peptide structures. BIV-TAR RNA is a retroviral regulatory element whose interaction with proteins can be modeled by peptides which bind the RNA. The peptide:RNA interaction includes an isoleucine which contributes to the free energy of association (Chen and Frankel 1995). In two NMR structures of the peptide:RNA complex (Puglisi et al. 1995; Ye et al. 1995) the isoleucine side chain abuts and makes its apparently stabilizing interaction with the hydrophobic H5-H6 edge (the side away from the WC pairing face) of a crucial U base (compare Sundquist 1996). In fact, the amino acid side-chain distinctions made by this site, deduced from binding measurements on substituted peptides (Chen and Frankel 1995), are of the order of those also measured in the selected RNA sites above. Notably, conserved U's are prominent in all three selected internal-loop sites for aliphatic amino acids.

G's in hydrophobic RNA sites may also be expected. The arginine site in Fig. 2d (Yang et al. 1996) contains a hydrophobic contact in which the aliphatic part of the arginine side chain is extended across the face of a guanine base. Such aliphatic-guanine base contacts are among the earliest-identified hydrophobic ribonucleotide base-amino acid interactions. They are evident in the GTP-binding pockets of EF-Tu (la Cour et al. 1985) and ras (Pai et al. 1989). Thus both conserved G's and G:U appositions in the three aliphatic sites in Fig. 3 may be explicable in terms of already-known interactions.

Furthermore, the small G/U motifs that occur in the sites selected for these two aliphatic amino acid side chains recur in greatly expanded form in nucleic acid sites for large hydrophobic ligands. A DNA oligomer selected to bind and introduce a divalent ion within the hydrophobic porphyrin ring system (Li and Sen 1996) is composed of 82% G and T (Fig. 4a). An RNA oligomer selected to catalyze a similar porphyrin metalation reaction (Conn et al. 1996) is also very G and U rich (Fig. 4b). In the latter case the active-site residues are known because they are completely conserved in independent



**Fig. 4.** Nucleic acid sites for large hydrophobic ligands contain many G's and U/T's. **a** Catalytic DNA that binds porphyrin (Li and Sen 1996). **b** Catalytic RNA that binds porphyrin (Conn et al. 1996). G's and U/T's are *boldfaced*.

isolates; 84% of the conserved loop nucleotides are G and U. Figure 4 shows these remarkable structures with G and U/T in boldface to make their prevalence evident. Thus larger hydrophobic sites can also be constructed using G and U, and conversely, conservation of G/U motifs in an RNA may indicate a hydrophobic site.

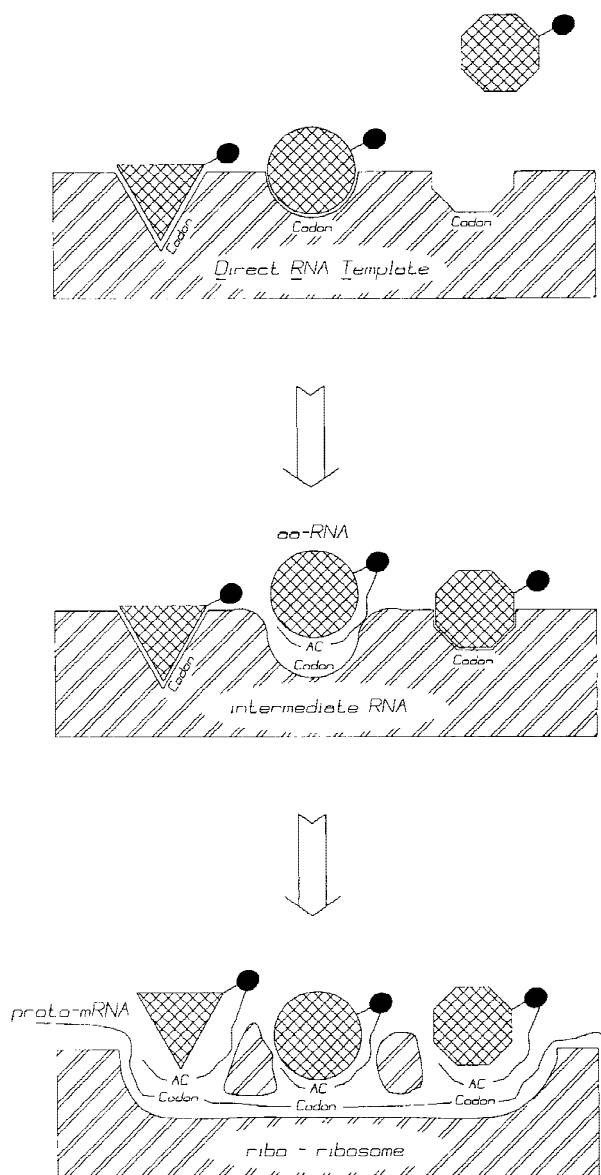
## A DRT Theory

I now consider the implications of RNA sites for amino acids for the evolution of the synthesis of peptides of predetermined sequence.

Simple RNA structures are capable of binding varied amino acids with substantial discrimination. This supports any scheme in which specific RNA:amino acid interactions underlie the genetic code. That is, the stereochemical theory championed by Woese et al. (1967) is strengthened, with respect to an alternative entirely arbitrary "frozen accident" (cf. Crick 1968). However, RNA is unexpectedly versatile, such that the code may be a frozen stereochemical accident. In other words, the code may preserve a set of biochemical interactions, but the choice of particular interactions now appears so broad that many other codes could have resulted.

A form of stereochemical theory which accommodates all the data above is shown in Fig. 5a, which defines the direct RNA templating hypothesis (DRT). For concision, I use DRT to refer to the RNA template, its action, and the hypothesis. The crucial notion is that specific peptides were first made by ordering carboxyl-activated amino acids using amino acid sites in an RNA template. Within the sites are subsequences which will be selected, as translation evolves, to become modern coding sequences. Ordered peptides themselves are envisioned as being of value in the RNA-based biosphere, so that there was selection for the origination, and then improvement, of peptide synthesis.

There are at least three compelling arguments for a DRT theory. First and most important, the chemistry



**Fig. 5.** A hypothesis for the origin and evolution of the genetic code. **a** Direct RNA templating (DRT) to specify peptide sequences. The small filled ovoids are carboxyl-activating groups (ribose-esterified aminoacyl-A, in the favored form of the hypothesis); the large cross-hatched shapes represent different amino acid side chains. Sequences designated “codons” within the binding sites are not codons at the time of panel a, but are so called to clarify their connection with later events. **b** Somewhat later—the appearance of aminoacyl-RNA (aa-RNA). AC, anticodon. **c** Still later—separation of the DRT into mRNA and protoribosome. Ribo-ribosome, a hypothetical RNA precursor of the nucleoprotein ribosome, a protoribosome.

required for its operation has been shown to exist. As pointed out above, amino acid sites specific to both polar and nonpolar amino acids can be folded from RNA. Binding constants are sufficient to secure the amino acids from relatively dilute solutions and specificities sufficient reproducibly to produce oligomers of some length. Sequences destined to be selected as coding triplets could in fact exist within the primordial acid sites because they demonstrably exist within RNA sites today

(Figs. 1–3). Such prospective template RNAs containing amino acid sites could have been relatively small; for example, they need be no larger than RNAs already made using activated nucleotides and mineral catalysis (Ferris et al. 1996). There is therefore a substantial body of experimental support for DRT.

Second, the proposed primordial DRT system is simple, consisting of only two elements: RNA templates and activated amino acids. Simplicity would be a cardinal virtue amidst the irreproducibility prevailing in a forming or primitive organisms. No simpler system is possible; but if the addition of a third element (e.g., a peptidyl transferase) is thought to be essential, it can likely be incorporated into the initial DRT, as modern RNAs may perform similar catalysis (Noller et al. 1992; Lohse and Szostak 1996; Welch et al. 1997).

Third, the DRT hypothesis meets the requirement of continuity (Orgel, 1968), which says that ancestral systems should give way smoothly to their modern counterparts without the need for discontinuities and ad hoc innovations. Figures 5b and c are not canonical parts of the DRT hypothesis but illustrate a plausible continuity.

In Fig. 5b, ancestral transfer RNAs arise to adapt amino acids to RNA binding sites on templates whose coding was still of a mixed type. At this stage, therefore, particular RNA triplets, which previously may have played quite variable roles in their amino acid sites, become coding sequences for the first time. Selection for specific amino acid sequences forces these coding sequences to become unique. If the scheme in the drawing is adopted, DRT specifies that the codons rather than the anticodons (e.g., Lacey et al. 1985) survive for ancestral binding sites, as suggested by the group I example. If the carboxyl activation of the amino acids is achieved by esterification to adenine nucleotide, then proto-tRNAs can arise by extension of the primordial activating group, as suggested in Fig. 5b. Ancient aminoacyl-RNAs are supported by the isolation of small modern RNA catalysts for synthesis of ribose 2'/(3')-esterified aa-RNA from ubiquitous biological reactants (Illangasekare et al. 1997).

The added RNA of the proto-tRNAs (Fig. 5b) would also likely make new functions possible, for example, allowing the same events to occur during translation for every aminoacyl-RNA, so that the peptide chain extension cycle could be standardized and optimized, ultimately giving rise to modern specialized ribosomal sites.

Messenger RNA (mRNA) is created when all peptide extension is performed using aminoacyl-RNAs, pairing through their anticodons (AC) to the collected portions of the original sites which have now come to function as their triplet codons (Fig. 5c). The residual (non-mRNA) part of the DRT, which may have acquired stimulatory activities, e.g., for aminoacyl-RNA binding, may survive as a protoribosome (the ribo-ribosome). This evolutionary transition from the DRT to mRNA would be facilitated

tated if a variety of RNA sites could offer their primordial codons in similar structural contexts, so that a similar pathway could progressively capture them all. Structural similarities among amino acid sites known today (above) suggest that such a conserved pathway from top to bottom in Fig. 5 might be realized using asymmetric internal RNA loops as the ancestral amino acid sites.

### Strong Versus Weak DRT

Discussion of further evidence for DRT requires distinction between the weak and the strong forms of the hypothesis.

“Restrained” or weak DRT asserts that the chemistry of RNA allows coding for peptides by the formation of RNA template surfaces having ordered amino acid binding sites. Accordingly, aboriginal peptide synthesis templates were RNA surfaces of some type. Many elements of the weak form (Yarus 1991) of the DRT hypothesis have already been demonstrated in the results just reviewed. The proof of plausibility is incomplete, as only a few amino acids have been investigated and no such set of ordered peptides has actually been synthesized to show that all conceivable chemical difficulties can be overcome. However, such a demonstration seems within reach.

The strong, or “exuberant,” form (Yarus 1991) of the DRT hypothesis is not only that RNA templates were ancestral to the modern translation apparatus, but that specific RNA:amino acid complexes can be identified as progenitors of the present genetic code. For example, strong DRT predicts that coding sequences occur at higher-than-expected frequencies in selected amino acid binding sites. However, such as experimental proof faces two related, reinforcing difficulties.

First, coding relies on triplets, which are intrinsically frequent sequences. Let us say that RNA amino acid sites consist of a certain number of nucleotides. The number must be small to be plausible, since amino acids are about a third the size of nucleotides, and therefore the space around a bound amino acid is small, even if several layers are contemplated. For illustration, assume 10 nucleotides “near” the bound amino acid. Then, for an amino acid having three codons, like isoleucine, the probability of finding at least one codon in a run of 10 randomized nucleotides (taken as contiguous for simplicity) is  $\approx 0.3$ . Thus the finding that both prevalent isoleucine sites among selected isoleucine-binding RNAs contain isoleucine triplets (Fig. 3b) does not allow one confidently to impute meaning to the observation. For arginine, with six codons, the statistical situation is even less favorable: in 10 randomized nucleotides with eight possible triplets, the probability is  $\approx 0.5$  of finding arginine triplets. Thus the finding above, that  $6/9 = 0.67$  of

known sites have such triplets, or that there is a mean of 1.3 triplets per site (rather than the  $\approx 0.75$  codon/site expected at random), cannot be interpreted as support for strong DRT.

Second, RNA has proven to be far more versatile than once imagined. As emphasized above, selections are conducted to isolate prevalent examples rather than to characterize the total population of sites. Each new method of selection finds new sequences, at least for arginine. Thus many sites remain to be discovered, each having the same *a priori* claim to primordial status as the ones we know. When looking for particular sequences, we can never know that we have looked correctly, or far enough.

Thus the need to consider both strong and weak DRT: experimental evidence of RNA versatility now available has opposite effects on the weak and strong forms of DRT. It strengthens the weak form, which seems at this point very plausible indeed. But while the triplet frequency evidence pertinent to strong DRT is consistent with the hypothesis (Figs. 1–3; see above), statistically compelling proof of strong DRT via an elevated frequency of coding sequences in selected amino acid sites seems out of reach.

### Beyond Statistical Considerations

But there are other possible kinds of evidence: strong DRT asserts that the group I active center is the progenitor of at least five of six modern arginine codons. This group I example (Fig. 1) demonstrates that discovery of additional potentially ancient associations between an amino acid and its coding triplets could strengthen strong DRT, potentially linking it to a known phylogeny.

In addition, strong DRT would be strengthened if it can explain, beyond the associations of individual amino acids, the evident general organization of the code. That such an explanation is possible is suggested by the conservation of U in sites for aliphatic side chains (see above) and the use of the H5–H6 edge of U as a hydrophobic RNA element in a side-chain binding site (Puglisi et al. 1995; Ye et al. 1995; Sundquist 1996). It has long been noted that U is the central nucleotide in the codons of hydrophobic amino acids. The leftmost column in the standard coding table (second-position U) contains UUY–phe, UUR–leu, CUN–leu, AUH–ile, AUG–met, and GUN–val. That this list includes the hydrophobic amino acids in Fig. 3 may not, then, be a coincidence.

*Acknowledgments.* Thanks are due Robin Gutell for the data shown in Fig. 1. I appreciate many suggestions from my group which clarified a draft manuscript. This project was supported by NIH Research Grants GM30881 and GM48080.



## References

- Burgstaller P, Kochoyan M, Famulok M (1995) Structural probing and damage selection of citrulline- and arginine-specific RNA aptamers identify base positions required for binding. *Nucleic Acids Res* 23:4769–4776
- Cedergren R, Miramontes P (1996) The puzzling origin of the genetic code. *TIBS* 21:199–200
- Chen L, Frankel AD (1995) A peptide interaction in the major groove of RNA resembles interactions in the minor groove of DNA. *Proc Natl Acad Sci USA* 92:5077–5081
- Conn MM, Prudent JR, Schultz PG (1996) Porphyrin metalation catalyzed by a small RNA molecule. *J Am Chem Soc* 118:7012–7013
- Connell GJ, Yarus M (1994) RNAs with dual specificity and dual RNAs with similar specificity. *Science* 264:1137–1141
- Connell GJ, Illangasekare M, Yarus M (1993) Three small ribooligonucleotides with specific arginine sites. *Biochemistry* 32:5497–5502
- Crick FHC (1968) The origin of the genetic code. *J Mol Biol* 38:367–379
- Ellington AD, Szostak JW (1990) *In vitro* selection of RNA molecules that bind specific ligands. *Nature* 346:818–822
- Famulok M (1994) Molecular recognition of amino acids by RNA-aptamers: an L-citrulline binding RNA motif and its evolution into an L-arginine binder. *J Am Chem Soc* 116:1698–1706
- Famulok M, Szostak W (1992) Stereospecific recognition of tryptophan agarose by *in vitro* selected RNA. *J Am Chem Soc* 114:3990–3991
- Ferris JP, Hill AR Jr, Liu R, Orgel LE (1996) Synthesis of long prebiotic oligomers on mineral surfaces. *Nature* 381:59–61
- Geiger A, Burgstaller P, Eltz Hvd, Roeder A, Famulok M (1996) RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. *Nucleic Acids Res* 24:1029–1036
- Golden BL, Cech TR (1996) Conformational switches involved in orchestrating the successive steps of group I RNA splicing. *Biochemistry* 35:3754–3763
- Harada K, Frankel AD (1995) Identification of two novel arginine binding DNAs. *EMBO J* 14:5798–5811
- Hicke BJ, Christian EL, Yarus M (1989) Stereoselective arginine binding is a phylogenetically conserved property of group I self-splicing RNAs. *EMBO J* 8:3843–3851
- Hirao I, Ellington AD (1995) Re-creating the RNA world. *Curr Biol* 5:1017–1022
- Illangasekare M, Kovalchuk O, Yarus M (1997) Essential structures of a self-aminoacylating RNA. *J Mol Biol* 274:519–529
- Kim JL, Nikolov DB, Burley SK (1993) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* 365:520–527
- Kim Y, Geiger JH, Hahn S, Sigler PB (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature* 365:512–520
- Lacey JC Jr, Hall LM, Mullins DW Jr (1985) Rationalization of some genetic anticodon assignments. *Origins Life* 16:69–79
- la Cour TF, Nyborg J, Thirup S, Clark BFC (1985) Structural details of the binding of GDP to elongation factor Tu from *E coli* as studied by X-ray crystallography. *EMBO J* 4:2385–2388
- Li Y, Sen D (1996) A catalytic DNA for porphyrin metallation. *Nature Struct Biol* 3:743–747
- Lohse PA, Szostak JW (1996) Ribozyme-catalyzed amino-acid transfer reactions. *Nature* 381:442–444
- Majerfeld I, Yarus M (1994) An RNA pocket for an aliphatic hydrophobe. *Nature Struct Biol* 1:287–292
- Majerfeld I, Yarus M (1998) Isoleucine binding sites with essential coding sequences. *RNA* 4:471–478
- Michel F, Hanna M, Green R, Bartel DP, Szostak JW (1989) The guanosine binding site of the *Tetrahymena* intron. *Nature* 342:391–395
- Noller HF, Hoffarth V, Zimmick L (1992) Unusual resistance of peptidyl transferase to protein extraction methods. *Science* 256:1416–1419
- Orgel LE (1968) Evolution of the genetic apparatus. *J Mol Biol* 38:381–393
- Pai EF, Kabsh W, Kregel U, Holmes KC, John J, Wittinghofer A (1989) Structure of the guanine nucleotide binding domain of the *Ha-ras* oncogene product p21 in the triphosphate conformation. *Nature* 341:209–214
- Puglisi JD, Chen L, Blanchard S, Frankel AD (1995) Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science* 270:1200–1203
- Robertson DL, Joyce GF (1990) Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* 344:467–468
- Shub DA (1991) The antiquity of group I introns. *Curr Opin Genet Dev* 1:478–484
- Sundquist WI (1996) Tattle tales. *Nature Struct Biol* 3:8–11
- Tan R, Frankel AD (1995) Structural variety of arginine-rich RNA-binding peptides. *Proc Natl Acad Sci USA* 92:5282–5286
- Tao J, Frankel AD (1996) Arginine-binding RNAs resembling TAR identified by *in vitro* selection. *Biochemistry* 35:2229–2238
- Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to T4 DNA polymerase. *Science* 249:505–510
- Welch M, Majerfeld I, Yarus M (1997) 23S rRNA similarity from selection for peptidyltransferase mimicry. *Biochemistry* 36:6614–6623
- Werner MH, Huth JR, Gronenborn AM, Clore GM (1995) Molecular basis of human 46X,Y sex reversal revealed from the 3-dimensional solution structure of the human SRY-DNA complex. *Cell* 81:705–714
- Woese CR, Dugre DH, Dugre AS, Kondo M, Saxinger WC (1967) On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symp Quant Biol* 31:723–736
- Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok M (1996) Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science* 272:1343–1347
- Yarus M (1988) A specific amino acid binding site composed of RNA. *Science* 240:1751–1758
- Yarus M (1991) An RNA-amino acid complex and the origin of the genetic code. *New Biol* 3:183–189
- Yarus M (1993) An RNA-amino acid affinity. In: Gesteland RF, Atkins JF (eds) *The RNA world*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 205–217
- Yarus M, Christian EL (1989) Genetic code origins. *Nature* 342:349–350
- Yarus M, Majerfeld I (1992) Co-optimization of ribozyme substrate stacking and L-arginine binding. *J Mol Biol* 225:945–949
- Yarus M, Illangasekare M, Christian EL (1991a) Selection of small molecules by the *Tetrahymena* catalytic center. *EMBO J* 19:1297–1304
- Yarus M, Illangasekare M, Christian EL (1991b) An axial binding site in the *Tetrahymena* precursor RNA. *J Mol Biol* 222:995–1012
- Ye X, Kumar RA, Patel DJ (1995) Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem Biol* 2:827–840
- Zinnen S, Yarus M (1995) An RNA pocket for the planar aromatic side chains of phenylalanine and tryptophan. *Nucleic Acids Symp Ser* 33:148–151